Performance evaluation of grid-enabled registration algorithms using bronze-standards

T. Glatard[1,2], X. Pennec[1], and J. Montagnat[2]

[1] INRIA Sophia - Projet Asclepios, 2004 Route des Lucioles BP 93
06902 Sophia Antipolis Cedex, France
{Xavier.Pennec,Tristan.Glatard}@sophia.inria.fr
[2] CNRS - I3S unit, RAINBOW team, 930 route des Colles, BP 145
06903 Sophia Antipolis Cedex, France
{glatard,johan}@i3s.unice.fr

Abstract. Evaluating registration algorithms is difficult due to the lack of gold standard in most clinical procedures. The *bronze standard* is a real-data based statistical method providing an alternative registration reference through a computationally intensive image database registration procedure. We propose in this paper an efficient implementation of this method through a grid-interfaced workflow enactor enabling the concurrent processing of hundreds of image registrations in a couple of hours only. The performances of two different grid infrastructures were compared. We computed the accuracy of 4 different rigid registration algorithms on longitudinal MRI images of brain tumors. Results showed an average subvoxel accuracy of 0.4 mm and 0.15 degrees in rotation.

1 Performance evaluation using bronze standards

The accuracy performances of registration algorithms are critical for many clinical procedures but quantifying them is difficult due to the lack of gold standard in most clinical applications. To analyze registration algorithms, one may consider them as black boxes that take images as input and output a transformation. The performance evaluation problem is to estimate the quality of the transformation. However, no registration algorithm will perform the same for all types of input data. For instance, one algorithm may perform very well for multimodal MR registration but poorly for SPECT/CT. This means that the evaluation data set has to be representative of the clinical application problem we are targeting: all sources of perturbation in the data should be represented, such as acquisition noise and artifacts, pathologies, etc, and that we cannot just conclude from one experiment that one algorithm is better than the others for all applications.

1.1 Performance quantifiers

As far as the registration result is concerned, one can distinguish between gross errors (convergence to wrong local minima) and small errors around the exact transformation. The *robustness* can be quantified by the size of the basin of attraction of the right solution or by the probability of false positives. The small errors may be sorted into *systematic biases*, *repeatability* and *accuracy* [1]. The repeatability accounts for the errors due to internal parameters of the algorithm, mainly the initial transformation, and to the finite numerical accuracy of the optimization algorithm, while the external error accounts for the propagation

of the data errors into the optimization result. It is important to notice that accuracy measures the error with respect to the truth (which may be unknown), while the precision or repeatability only measures the deviation from the average value, i.e. it does not take into account systematic biases, which are often hidden. For instance, a calibration error in the acquisition system will consistently bias all the images acquired with that device. Unless another calibration is done or an external reference is used (e.g. another acquisition device), there is no way to detect such a bias. In terms of statistical modeling, this means that all the potential error sources should be made random in order to be included.

In a statistical setting, considering the input data and the output transformation as random variables naturally leads to quantify the precision (resp. accuracy) of the transformation as the standard deviation or expected RMS distance to the mean (resp. the exact) transformation, or more interestingly with the covariance matrix as the transformation uncertainty is usually non isotropic (e.g. radians and millimeters for rotation and translation part of a rigid transformation). Then, the variability of the transformation can be propagated to some target points using standard first order linearizations to obtain the covariance on the transformed test points, or its trace, the variance (see e.g. [6]).

1.2 Performance evaluation

One of the simplest evaluation schemes is to simulate noisy data and to measure how far is the registration result from the true one (the ground truth is obviously known). The main drawback of synthetic data is that it is very difficult to identify and model faithfully all the sources of variability, and especially unexpected events (pathologies, artifacts, etc). Forgetting one single source of error (e.g. camera calibration errors in 2D-3D registration) automatically leads to underestimation of the final transformation variability. In some cases, however, images may be faithfully simulated (e.g SPECT and MRI), with a very high computational cost due to the complexity of image acquisition physics.

The second evaluation level is to use real data in a controlled environment, for instance imaging a physical phantom. There is possibly a gold standard, if one can precisely measure the motion or deformation of the phantom with an external apparatus. However, it is difficult to test all the clinical conditions (e.g. different types or localizations of pathologies). Moreover, it is often argued that these phantoms are not representative of real in vivo biological systems. One level closer to the reality, experiments on cadavers correctly take into account the anatomy, but fail to exhibit all the errors due to the physiology. Moreover, images may be very different from the in-vivo ones.

We tackle in this paper the last level of evaluation methods, which relies on a database of in-vivo real images representative of the clinical application. Such a database can be large enough to span all sources of variability, but there is usually no gold standard registration to compare with. One method is to perform a cross comparison of the criteria optimized by different algorithms [2]. However, this does not give any insight about the transformation itself. A more interesting method for registration evaluation is the use of consistency loops [3, 4]. The principle is to compose transformations that form a closed circuit and to measure the difference of the composition from the identity. This criterion does

not require any ground truth, but it only measures the repeatability as any bias will get unnoticed. A last type of methods is to see the ground truth as a hidden variable, and to estimate concurrently the ground truth and the quality as the distance of our results to this reference (EM like algorithms). This method was exemplified for the validation of segmentation by the STAPLE algorithm [5].

1.3 The Bronze Standard method

The principle of the bronze standard method is similar but concerns registration: from a set of registrations between images, we want to estimate the exact transformations, and the variability of the registration results with respect to these references. Let us assume that we have n images of the same organ of the patient and m methods to register them, i.e. $m \times n^2$ transformations $T_{i,j}^k$ (we denote here by k the index of the method and by i and j the indexes of the reference and target images). Our goal here is to estimate the $n-1$ free transformations $\bar{T}_{i,i+1}$ that relate successive images and that best explain the measurements $T_{i,j}^k$.

The bronze standard transformation between images i and j is obtained by composition: $\bar{T}_{i,j} = \bar{T}_{i,i+1} \circ \bar{T}_{i+1,i+2} \circ \ldots \circ \bar{T}_{j-1,j}$ if $i < j$ (or the inverse of both terms if $j > i$). The free transformation parameters are computed by minimizing the prediction error on the observed registrations:

$$C(\bar{T}_{1,2}, \bar{T}_{2,3}, \ldots, \bar{T}_{n-1,n}) = \sum_{i,j \in [1,n], k \in [1,m]} d\left(T_{i,j}^k, \bar{T}_{i,j}\right)^2 \tag{1}$$

Here, d is a distance function between transformations chosen as a robust variant of the left invariant distance on rigid transformation developed in [6]:

$$d(T_1, T_2) = \min\left(\mu^2(T_1^{(-1)} \circ T_2)\ ,\ \chi^2\right) \quad \text{with} \quad \mu^2(R(\theta, n), t) = \theta^2/\sigma_{\mathrm{r}}^2 + \|t\|^2/\sigma_{\mathrm{t}}^2$$

where θ is the angle of rotation R and n is the unitary vector defining its axis. t is the translation vector of the transformation. Details on the general methods for doing statistics on Riemannian manifolds and Lie groups are given in [7].

In this process, we do not only estimate the optimal transformations, but also the rotational and translational variance of the "transformation measurements", which are propagated through the criterion to give an estimate of the variance of the optimal transformations. Of course, these variances should be considered as a fixed effect (i.e. these parameters are common to all patients for a given image registration problem, contrarily to the transformations) so that they can be computed more faithfully by multiplying the number of patients.

The estimation $\bar{T}_{i,i+1}$ is called *bronze standard* because the result converges toward the perfect registration as the number of methods m and the number of images n increases. Indeed, considering a given registration method, the variability due to the noise in the data decreases as the number of images n increases, and the registration computed converges toward the perfect registration up to the intrinsic bias introduced by the method. Now, using different registration procedures based on different methods, the intrinsic bias of each method also becomes a random variable, which is hopefully centered around zero and averaged out in the minimization procedure. The different bias of the methods are now integrated into the transformation variability. To fully reach this goal, it is important to use as many independent registration methods as possible.

Criterion (1) is in fact the log-likelihood of the observations $T_{i,j}^k$ assuming Gaussian errors around the bronze standard registrations with a variance σ_r^2 on the rotation and σ_t^2 on the translation. An important variant is to relax the assumption of the same variances for all algorithms, and to unbias their estimation. This can be realized by using only $m-1$ out of the m methods to determine the bronze standard registration, and use the obtained reference to determine the accuracy of the last method (a kind of leave-one-method-out test).

2 Gridifying registration algorithms

The large amount of input data and registration algorithms required to compute the bronze standard makes this method very compute intensive. A grid infrastructure can handle the load of the computations involved and help in managing the medical image database to process. A grid is a pool of shared computing and storage resources accessible through a middleware software layer which hides as much as possible the complexity of the infrastructure to the user. Those platforms are designed to help users to share and execute efficiently their algorithms and data, which fulfills the needs of the bronze standard application.

2.1 Interoperability to compare and share algorithms

Sharing registration algorithms implies that each registration service is semantically described. A complete system would include an ontology of registration problems (modalities, anatomical region, rigid or non-rigid registration, etc), of registration algorithms (input/output data types, method used, etc) and of image and transformation formats. In our case, converting rigid transformations formats was quite simple (although some transformations were expressed in different bases) but extending that to non-rigid transformations is still open.

From a practical point of view, we wrapped each registration algorithm into a standard Web service. Those Web services are responsible for the grid execution of their algorithm on the data sets specified at invocation. Algorithms are thus

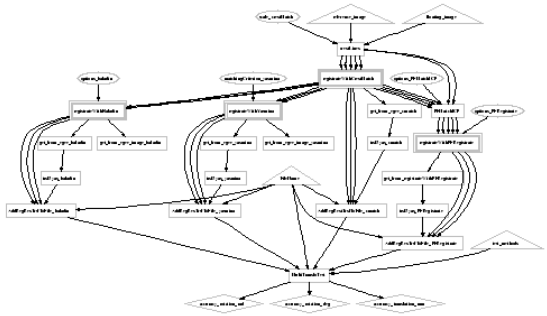

Fig. 1. Bronze standard workflow. Each double squared box represents a registration algorithm. Lightweight computing tasks such as data transfers and format transformations are represented with simply-squared boxes and arrows show computation dependencies. Triangles figure the inputs of the workflow, rhombs the outputs and ellipses the parameters. The final box is the bronze standard computation.

standard black boxes, ready to be assembled into an application. To minimize the user effort for the gridification of its algorithm, we developed a generic wrapper that is able to submit a job on the grid, given a simple description of the corresponding command line.

2.2 Workflow description and execution

Once registration algorithms are wrapped, one can describe the global bronze standard application. Data links are first specified between outputs and inputs of algorithms, in order to define the data pipeline. Control links may also be specified in order to describe precedence constraints between algorithms. We chose to describe the workflow with the Scufl language [8] which is a good trade off between high expressiveness and simplicity. The obtained bronze-standard workflow is depicted on figure 1 and fully described in Section 3.

To efficiently execute such a workflow on a grid, we developed a workflow manager called MOTEUR [9]. It particularly focuses on optimizing the time performances, which are critical in the case of data-intensive applications such as our bronze standard. MOTEUR enables three different kinds of parallelism (workflow, data and service parallelism), in order to exploit the massively parallel resources available on the grid infrastructure. Moreover, it groups sequential jobs to lower the number of services invocations and minimize the grid overhead resulting from jobs submission, scheduling and data transfers. Finally, MOTEUR can execute workflows on grid systems with very different scales. We made experiments on the EGEE production grid[3] (including 18,000 CPUs all over the world and 5 PB storage capacity), as well as on the Grid5000 experimental infrastructure[4] (2,000 CPUs and hundreds of GB storage capacity).

3 Experiments

We are targeting the clinical follow-up of the radiotherapy of brain tumors, which requires several registrations. To optimize the dose planning, a deformable atlas to patient registration is performed to segment target volumes and organs at risk. To be more accurate, multimodal images are often co-registered. Last but not least, the tumor evolution and the result of the treatment are assessed in follow-up images, thanks to a monomodal rigid registration. This is the registration problem that we consider in this paper. Quantifying its accuracy is important to ensure the precision of the tumor evolution estimation in the assessment of the efficiency of clinical treatments. Precisely registered longitudinal studies may be used to validate the quality (reproducibility and accuracy) of segmentation algorithms used for radiotherapy planning.

To evaluate all these registration / segmentation problems, a database of 110 patients with 1 to 6 times points and MR T2, T1 and gadolinium injected T1 modalities was acquired at a local cancer treatment center (courtesy of Dr Pierre-Yves Bondiau from the "Centre Antoine Lacassagne", Nice, France) on a Genesis Signa MR scanner. Among them, 29 have more than one time point and were suitable to inclusion in our rigid registration evaluation study. We chose to select only the injected T1 images in a first step. These images are more demanding for registration than other MRI sequences as the gadolinium uptake

[3] Enabling Grids for E-sciencE, http://www.eu-egee.org
[4] Grid5000 national grid, http://www.grid5000.org

is likely to vary at different time points, leading to local intensity outliers. All T1i images are 256×256×60 ×16 bits.

We considered four different registration algorithms. Two of them are intensity-based: `Baladin` [10] has a block matching strategy optimizing the coefficient of correlation and a robust least-trimmed-squares transformation estimation; `Yasmina` uses the Powell algorithm to optimize the SSD or a robust variant of the correlation ratio (CR) [4]. The two others are feature-based and match specific points crest lines with different strategies [11]: `CrestMatch` is a prediction-verification method and `PFRegister` is an ICP algorithm extended to features more complex than points. In the computation of the bronze standard registration, `CrestMatch` is used to initialize all the other algorithms close to the right registration. This allows us to ensure that all algorithms converge toward the same (hopefully global) minimum. A visual inspection is performed a posteriori on the bronze standard registration to ensure that this "optimal" transformation is indeed correct. As we are focusing on accuracy and not on the robustness, this initialization does not bias the evaluation. Figure 1 illustrates the application workflow.

3.1 Accuracy results

The workflow was run on the 29 selected patients with $\sigma_r = 0.15$ degrees, $\sigma_t = 0.42$ mm and a χ^2 value of 30. A high number of registration results were rejected in the robust estimation of the bronze standard transformations. A visual inspection revealed that there was a scaling and shear problem in the yz plane for one of the image involved in each of these rejected registrations. A detailed analysis of the DICOM headers showed that the normal to the slices (xy plane), given by the cross product of the `Image Orientation` vectors, was not perfectly parallel to the slice trajectory during the acquisition (axis obtained from the `Image Position` field). This tilt was found to be +1.19 degree in most of the images and -1.22 degree in 13 images. It seems that nothing in the DICOM standard ensures that 3D images are acquired on an orthogonal grid: it would be interesting to better specify the acquisition protocols on the MR workstation (the radiologists were even not aware of that tilt!).

Thus, images are not in an orthogonal coordinate system and should be either registered with an affine transformation (which adds 6 additional parameters among which only one -the tilt- has a physical justification) or the tilt should be taken into account within the rigid registration algorithm, but this solution was not implemented for the algorithms we were considering. As the tilt was small, we chose not to resample the images (in order to keep the original image quality), but rather to perform an uncorrected rigid registration within the group of images with a positive tilt only. This led us to remove 13 images among the 82, and 4 patients for which only one image was remaining (the statistics on the remaining number of patients, images and registrations are given in table 1).

The bronze standard workflow was run again with the same parameters on this reduced database of 25 patients. This time, only 20 registrations were rejected, among which 15 were concerning two patients with a very high deformation in the tumor area, leading to some global deformations of the brain (Fig. 2).

Number of time points:	2	3	4	6
Registration per patient (and per algorithm):	2	6	12	30
Patients (including/without tilted images):	15/**15**	6/**7**	7/**2**	1/**1**
Total number of registrations:	120/**120**	144/**168**	336/**96**	120/**120**

Table 1. Summary statistics about the image database used.

Fig. 2. Example of a slice of two registered images with a high deformation.

In that situation the rigid motion assumption does not hold any more and several "optimal" rigid registration may be valid depending on the area of the brain. The last 5 rejected transformations involve two acquisitions with phase-encoded motion artifacts which impacted differently feature-based and intensity-based registration algorithms, leading to two non-compatible sets of transformations. However, it was not possible to visually decide which result was the "right" one.

Excluding these 20 transformations which correspond to special conditions where the rigid assumption does not really hold, we obtained mean errors of 0.130 degree on the rotations and 0.345 mm on the translations. The propagation of this error on the estimated bronze standard leads to an accuracy of 0.05 degree and 0.148 mm. We then determined the unbiased accuracy of each of the 4 algorithms by comparing its results to the bronze standard computed from the 3 others methods. Results are presented in table 2 and show slightly higher but equivalent values for all algorithms.

Algorithm	$\sigma_r(deg)$	$\sigma_t(mm)$
CrestMatch	0.150	0.424
PFRegister	0.180	0.416
Baladin	0.139	0.395
Yasmina	0.137	0.445

Table 2. Accuracy results

Image pairs	12	66	126
Sequential	2h40min	14h40min	28h
Grid5000	10min	35min	2h10min
EGEE	2h10min	3h22min	4h57min
Grid5000 speed-up w.r.t EGEE	13	5.8	2.3

Table 3. Execution times

3.2 Grid-computing results

The execution times of the whole workflow was compared on the EGEE and Grid5000 (Sophia shared cluster of 105 nodes) platforms and on the sequential case, for different numbers of image pairs to register (Table 3). Even though the EGEE production infrastructure gathers many more processors than the Grid5000 cluster, the workflow was always faster on the Grid5000 cluster. This is explained by the high overhead introduced by the EGEE grid, coming from the large scale of this platform and its multi-users nature. However, the speed-up obtained on the Grid5000 cluster vs EGEE is decreasing with the number of input images. The Grid5000 cluster progressively enters a saturation phase, where all the available processors are used by the application, while the EGEE grid is more scalable and less impacted by the growth of the input data set size.

4 Discussion

We propose in this paper a bronze standard evaluation framework to analyze the accuracy of rigid registration algorithms wrapped into web services, and a workflow engine to efficiently deploy this application on grids. The gridification of the application is motivated by the fact that both databases and registration algorithms may be more efficiently shared on grids. This is fundamental for bronze standards methods since the registration performance is converging from precision to accuracy only for a large number of data and algorithms.

Experiments demonstrate that the bronze standard method can be precise enough to detect very small deviations from the rigidity assumption (shears of 2 degrees) in images, and that the 4 rigid registration algorithms used actually reach a subvoxel accuracy of 0.15 degree in rotation and 0.4 mm in translation for the registration of longitudinal T1 injected 1x1x2mm images of the brain. Concerning the grid computing part, our results showed that the workflow engine was quite general and powerful. Moreover, execution times revealed that choosing the platform with the highest number of processors is not always the best solution as strong latencies may slow down the execution on wide infrastructures. The targeted grid infrastructure should thus be chosen according to the size of the problem to be solved.

References

1. P. Jannin et al. Validation of medical image processing in image-guided therapy. *IEEE TMI*, 21(12):1445–1449, December 2002.
2. P Hellier et al. Retrospective evaluation of intersubject brain registration. *IEEE Trans Med Imaging*, 22(9):1120–30, September 2003.
3. M. Holden et al. Voxel similarity measures for 3D serial MR brain image registration. *IEEE Trans. Med. Imaging*, 19(2):94–102, 2000.
4. A. Roche et al. Rigid registration of 3D ultrasound with MR images: a new approach combining intensity and gradient information. *IEEE TMI* 20(10):1038–1049, 2001.
5. SK Warfield et al. Simultaneous truth and perf. level estimation (staple): an algorithm for the validation of image segmentation. *IEEE TMI*, 23(7):903–921, 2004.
6. X. Pennec et al. Feature-based Registration of Medical Images: Estimation and Validation of the Pose Accuracy *MICCAI'1998*, LNCS 1496:1107–1114, 1998.
7. X. Pennec et al. Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements J. of Math. Imaging and Vision, 2006, to appear.
8. I. Oinn et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics journal*, 17(20):3045–3054, 2004.
9. T. Glatard et al. Efficient services composition for grid-enabled data-intensive applications. *Proc. of HPDC'06* 333–334
10. S. Ourselin et al. Block matching: A general framework to improve robustness of rigid registration. In *Proc. of MICCAI'2000*, LNCS 1935:557–566, 2000.
11. X. Pennec et al. Landmark-based registration using differential geometric features. *Handbook of Medical Imaging*, Chap. 31:499–513. Acad. Press, 2000.